REPORT RESUMES

ED 014 001 24 AA 000 269

SOCIAL SCIENCE EDUCATION CONSORTIUM. PUBLICATION 110, THE METHODOLOGY OF EVALUATION.

BY- SCRIVEN, MICHAEL

PURDUE UNIV., LAFAYETTE, IND.

REPORT NUMBER SSEC-PUB-110 PUB DATE MAR 66

REPORT NUMBER BR-5-0619-PUB-110

EDRS PRICE MF-$0.[redacted] HC-$[redacted] [redacted]

50 5.00 123 P.

DESCRIPTORS- *SOCIAL SCIENCES, *EVALUATION METHODS, *CURRICULUM EVALUATION, *TEACHER EVALUATION, *INSTRUCTIONAL MATERIALS, SOCIAL SCIENCE EDUCATION CONSORTIUM

THE AIM OF THIS PAPER IS TO EXHIBIT SOME OF THE PHILOSOPHICAL AND PRACTICAL DEFICIENCIES OF CURRENT CONCEPTIONS OF HOW EDUCATIONAL INSTRUMENTS SHOULD BE EVALUATED, AND TO SHOW WAYS FOR REDUCING THESE DEFICIENCIES. THE TERM "EDUCATIONAL INSTRUMENTS" IS USED TO INCLUDE SUCH THINGS AS NEW CURRICULUMS, PROGRAMED TEXTS, INDUCTIVE METHODS, AND INDIVIDUAL TEACHERS. THE MAIN FOCUS OF THE PAPER IS ON CURRICULUM EVALUATION, BUT IN THE AUTHOR'S OPINION, ALMOST ALL THE POINTS MADE TRANSFER IMMEDIATELY TO OTHER KINDS OF EVALUATION. SECTION HEADINGS ARE AS FOLLOWS--(1) OUTLINE, (2) GOALS OF EVALUATION VERSUS ROLES OF EVALUATION, (3) ARGUMENTS FOR AND AGAINST FORMATIVE AND SUMMATIVE EVALUATION, (4) EVALUATION VERSUS PROCESS STUDIES, (5) EVALUATION VERSUS ESTIMATION OF GOAL ACHIEVEMENT, (6) INSTRUMENTAL VERSUS CONSEQUENTIAL EVALUATION, (7) COMPARATIVE VERSUS NONCOMPARATIVE EVALUATION, (8) COMPARATIVE EVALUATION--THE CRITERIA OF EDUCATIONAL ACHIEVEMENT, (9) VALUES AND COSTS, (10) ANOTHER KIND OF EVALUATION--"EXPLANATORY EVALUATION," AND (11) CONCLUSIONS. THE DISCUSSION WHICH IS RELATIVELY ELEMENTARY AND ETIOLOGICAL IN THE EARLY SECTIONS PROGRESSES TO AN OCCASIONALLY MORE DIFFICULT AND GENERALLY MORE PRACTICAL LEVEL IN LATER SECTIONS. THIS PAPER WAS WRITTEN AS PART OF THE SOCIAL SCIENCE EDUCATION CONSORTIUM, A CURRICULUM PROJECT DESIGNED TO OUTLINE THE CONCEPTS, METHODS, AND STRUCTURE OF SEVERAL OF THE SOCIAL SCIENCES FOR USE BY TEACHERS AND CURRICULUM WORKERS AT ALL GRADE LEVELS. (JH)

DOCUMENT FILMED FROM BEST AVAILABLE COPY

BR 5-0619
PA 24

014001

SOCIAL SCIENCE EDUCATION CONSORTIUM

AA000269

SS
EC

THE METHODOLOGY
OF EVALUATION

Michael Scriven
Indiana University

Publication #110 of the
Social Science Education Consortium

Irving Morrissett, Executive Director
Purdue University, Lafayette, Indiana

The research reported herein was performed pursuant to a contract with the United States Department of Health, Education and Welfare, Office of Education, under the provisions of the Cooperative Research Program.

THE METHODOLOGY OF EVALUATION

0. Introduction.

Current conceptions of the evaluation of educational instruments (e.g. new curricula, programmed texts, inductive methods, individual teachers) are still inadequate both philosophically and practically. This paper attempts to exhibit and reduce some of the deficiencies. Intellectual progress is possible only because newcomers can stand on the shoulders of giants. This feat is often confused with treading on their toes, particularly but not only by the newcomer. I confess a special obligation to Professor Cronbach's work[1], and to valuable discussions with the personnel of CIRCE at the University of Illinois.

1. Outline.

The main focus of this paper is on curricular evaluation but almost all the points made transfer immediately to other kinds of evaluation. Section headings are reasonably self-explanatory and occur in the following order:

1. Outline.
2. Goals of Evaluation versus Roles of Evaluation.
3. Arguments for and against Formative and Summative Evaluation.

---

[1] "Evaluation for Course Improvement", Teachers' College Record, Vol. 64, No. 8, May 1963, reprinted in New Curricula (Ed. R. Heath, Pub. Harper & Rowe 1964, pp. 231-248); references in this paper are to the latter version.

4. Evaluation versus Process Studies.
5. Evaluation versus Estimation of Goal Achievement.
6. Instrumental versus Consequential Evaluation.
7. Comparative versus Non-Comparative Evaluation.
8. Comparative Evaluation - The Criteria of Educational Achievement.
9. Values and Costs.
10. Another Kind of Evaluation - 'Explanatory Evaluation'.
11. Conclusions.

The discussion in the earlier sections is relatively elementary and etiological, progressing to an occasionally more difficult and generally more practical level in later sections.

2. Goals of Evaluation versus Roles of Evaluation.

The aims of evaluation may be thought of in two ways. At the general level, we may talk of the goals of evaluation; in a particular educational context, of the roles of evaluation.

In general, we may say that evaluation attempts to answer certain types of question about certain entities. The types of question include questions of the form "How well does this instrument perform (with respect to such-and-such criteria)?", "Does it perform better than this other instrument?", "What does this instrument do (i.e. what variables from the group in which we are interested are significantly affected by its application)?", "Is the use of this instrument worth what it's costing?". Evaluation is itself a logical activity which is essentially similar whether we are trying to evaluate coffee machines or teaching machines, plans for a house or plans for a curriculum. The activity consists simply in the gathering and combining of performance data with a weighted set of goal scales to yield either comparative or numerical ratings.

But the role which evaluation has in a particular educational context may be enormously various; it may form part of a teacher training activity, of the process of curriculum development, of a field experiment connected with the improvement of learning theory, of an investigation preliminary to a decision about purchase or rejection of materials, it may be a data-gathering activity for supporting a request for tax increases or research support, or a preliminary to the reward or punishment of people as in an executive training program, a prison, or a classroom. Failure to make this rather obvious distinction between the roles and goals of evaluation, not necessarily in this terminology, is one of the factors that has led to the dilution of the process of evaluation to the point where it can no longer serve as a basis for answering the questions which are its goal. This dilution has sacrificed goals to roles. One can only be against evaluation if one can show that it is improper to seek for an answer to questions of the above kind, and this involves showing that there are no legitimate activities (roles) in which these questions can be raised, an extraordinary claim. Obviously the fact that evaluation is sometimes used in an inappropriate role hardly justifies the conclusion that we never need to know the answers to the goal questions.

One role that has often and sensibly been assigned to evaluation is as an important part of the process of curriculum development. Obviously such a role does not preclude evaluation of the final product of this process. Evaluation can obviously play several roles. Yet it is clear from the treatment of evaluation in some of the recent literature and in a number of recent research proposals involving several million dollars that the assumption is being made that one's obligations in the direction of evaluation are fully discharged by having it appear somewhere in a project. Not only can it have several roles with respect to one educational enterprise, but with respect to each of these it may have several goals. Thus, it may have a role in the improvement of the curriculum and with respect to this role several types of question (goals) may be raised, such as "Is the curriculum at this point really getting across the distinction between prejudice and commitment?", "Is it taking too large a proportion of the available time to make this point?", etc. In another role, the evaluation process may be brought to bear on the question of whether the entire finished curriculum, refined by use of the evaluation process in its first role, represents a sufficiently significant advance on the available alternatives to justify the expense of adoption by a school system.

One of the reasons for the tolerance or indeed encouragement of the confusion between roles and goals is the well-meaning attempt to allay the anxiety on the part of teachers that the word "evaluation" precipitates. By stressing the constructive part evaluation may play in non-threatening activities (roles) we slur over the fact that its goals are always the same - the estimation of merit, worth, value, etc. which all too clearly serves in another role as part of the evaluation of personnel and courses. It is unfortunate that we should be tackling anxiety about evaluation by reducing its importance and confusing its presentation; the loss in efficiency is too great. Business firms can't keep executives or factories on when they know they are not doing good work and a society shouldn't have to retain textbooks, courses, teachers and superintendents that do a poor job when a better performance is possible. The appropriate way to handle anxiety of this kind is by finding tasks for which a better prognosis is possible for the individual in question. Failure to evaluate pupils' performance leads to the gross inefficiencies of the age-graded classroom, and failure to evaluate teachers' performances leads to the correlative inefficiency of incompetent instruction. A little toughening of the moral fibre is required if we are not to shirk the social responsibilities of the educational branch of our culture. Thus, it may even be true that "the greatest service evaluation can perform is to identify aspects of the course where revision is desirable" (Cronbach, p.236), though it is not clear how one would establish this, but it is certainly also true that there are other extremely important services which must be done for almost any given project. And there are many contexts in which calling an evaluator in to perform a final evaluation of the project or person is an act of proper recognition of responsibility to the person, product or taxpayers. It therefore seems a little excessive to refer to this as simply "a menial role", as Cronbach does. It is obviously a great service if this kind of terminal evaluation (we might call it _summative_ as opposed to _formative_ evaluation) can demonstrate that a very expensive textbook is not significantly better than the competition, or that it is enormously better than any competitor. In more general terms it may be possible to demonstrate that a certain type of approach to e.g. mathematics is not yielding significantly better pupil performance on any dimension that mathematicians are prepared to regard as important. This would certainly save a great deal of expenditure of time and money and constitute a valuable contribution to educational development, as would the converse, favorable, result. Thus there seem to be a number of qualifications that would have to be made before one could accept a statement asserting the greater importance of formative evaluation by comparison with summative. ("Evaluation, used to improve the course while it is still fluid, contributes more to improvement of education than evaluation used to appraise a product already placed on the market." Cronbach, p.236) Fortunately we do not have to make this choice. Educational projects, particularly curricular ones, clearly must attempt to make best use of evaluation in both these roles.

Now any curriculum reformer is automatically engaged in formative evaluation, except on a very strict interpretation of 'evaluation'. He is presumably doing what he is doing because he judges that the material being presented in the existing curriculum is unsatisfactory. So as he proceeds to contruct the new material he is constantly evaluating his own material as better than that which is already current. Unless entirely ignorant of his shortcomings. as a judge of his own work, he is presumably engaged in field-testing the work while it is being developed, and in so doing he gets feedback on the basis of which he again produces revisions; this is of course formative evaluation. He is usually involved with colleagues, e.g. the classroom teacher or peers, who comment on the material as they see it - again, this is evaluation and it produces changes which are allegedly for the better. If the recommendation for formative evaluation has any content at all, it presumably amounts to the suggestion that a professional evaluator should be added to the curriculum construction project. There certainly can be advantages in this, but it is equally clear from practical experience that there can be disadvantages. But this argument is clearly not the same as the argument about summative evaluation. We devote part of the next section to a discussion of the pros and cons of formative evaluation.

3. Arguments for and against Formative and Summative Evaluation.

The basic fact is that the evaluator, while a professional in his own field, is usually not a professional in the field relevant to the curriculum being reformed or, if he is, he is not committed to the particular development being undertaken. This leads to clashes and failures to communicate of a kind which are all too familiar to project directors today.

From these 'failures of communication' between evaluators and teachers or curriculum makers there have sprung some unfortunate overreactions. The total anti-evaluation line is all too frequently a rationalization of the anxiety provoked by the presence of an external judge, not identified with or committed to (or perhaps even understanding) the ideals of the project. The equally indefensible opposite extreme is represented by the self-perceived tough-minded operationalist evaluator, all too likely to say "If you can't tell me what variables you are affecting, in operational terms, they can't be tested, and as long as they haven't been tested you haven't any reason for thinking you are making a contribution".

In order to develop a fair treatment of these views let us consider the difference between a contemporary educational project involving the development of a new curriculum or teaching method, and the co-authoring of a new ninth-grade algebra text by two or three teachers in the late 1930's. In the first place, the present projects are typically supported from government funds on a very large scale. The justification of this expenditure calls for some kind of objective evidence that the product was valuable. Moreover future support for work in this area or by these same workers requires some objective evidence as to their merit at this kind of job. Since there are not sufficient funds to support all applicants, judgements of comparative merit are necessary; and objective bases for this are trivially superior to mere person-endorsements by peers, etc. Finally, the enormous costs involved in the adoption of such products by school systems commit another great slice of taxpayers' money and this kind of commitment should presumably be made only on the basis of rather substantial evidence for its justification. In this context, summative evaluation is an inescapable obligation on the project director, and an obvious requirement by the sponsoring agency, and a desideratum as far as the schools are concerned. And since formative evaluation is part of a rational approach to producing good results on the summative evaluation, it can hardly be wholly eschewed; indeed, as we have shown, its occurrence is to some degree guaranteed by the nature of the case. But the separate question of whether professional evaluators should be employed depends very much upon the extent to which they do more harm than good - and there are a number of ways in which they can do harm.

They may simply exude a kind of skeptical spirit that dampens the creative fires of a productive group. They may be sympathetic but impose such crushing demands on operational formulation of goals as to divert too much time to an essentially secondary activity. ('Secondary' in the sense that there cannot be any evaluation without a curriculum.) The major compromise that must be effected is to have the evaluator recognise it as partly his responsibility to uncover and formulate a testable set of criteria for the course. He may be substantially helped by the fact that the project has explicitly espoused certain goals, or rejected others, and he will certainly be aided by their criticism of his formulations. However, the exchange has to be a two-way one; curriculum writers are by no means infallible, and often extremely prejudiced in describing their actual tendencies. Evaluators, on the other hand, are handicapped so long as they are less than fully familiar with the subject matter being restructured, and less than fully sympathetic with the aims of the creative group. Yet once they become identified with those aims, emotionally as well as economically, they lose something of great importance to an objective evaluation - their independence. For this reason the formative evaluators should be very sharply distinguished from the summative evaluators, with whom they may certainly work in developing an acceptable summative evaluation schema, but they should of course exclude themselves from any judgemental role.

There are other problems about the intrusion of evaluation into education, and the intrusion of an evaluator into the curriculum-making process. Several of these have been admirably expressed by J. Myron Atkin.[1] Some of them are taken up elsewhere in this paper, but some mention of two of them should be made here. The first suggestion is that testing for learning of certain rather delicate and pervasive concepts may be itself destructive, in that it makes the student too self-conscious about the role of a concept at too early a stage, thereby preventing its natural and proper development. The problem is that with respect to some of these concepts, e.g. symmetry, equilibrium and randomness, it might be the case that very little accretion occurs in the understanding of a child during any particular course or indeed any particular year of his education, but that tiny accretion may be of very great importance in the development of good scientific understanding. It would not show up on tests, indeed it might be stultified by the intrusion of tests, in any given year, but it has to be in the curriculum in order to produce the finished product that we desire. In this case, evaluation seems to be both incompetent and possibly destructive.

Such a possibility should serve as an interesting challenge to the creative curriculum-maker. While not dismissing it, he would normally respond by attempting to treat it more explicitly, perhaps at a somewhat later stage in the curriculum than it is normally first mentioned, and see whether some significant and satisfactory accretion of comprehension cannot be produced by this direct attack. Only if this failed would he turn to the evaluator and demand a considerably more sensitive instrument. Again, it would also be possible to deliberately avoid testing for this during all the early years of its peripheral introduction, and test only in the senior year in high school, for example. We can acknowledge the possibility that concerns Atkin and allow some extra material in the curriculum to handle it even without any justification from the early feedback from tests. Errors of excess are much less significant than errors of commission or omission, in curriculum-making.

---

[1] "Some Evaluation Problems in a Course Content Improvement Project", Journal of Research in Science Teaching, Vol. I, pp. 129-132 (1963).

Just as there are dangers from having a curriculum-making group discuss the present curriculum with teachers who are experienced in its use - although there are also possible advantages from this - so there are dangers and advantages in bringing the evaluator in too early. In such situations, some ingenuity on the part of the project director will often make the best of both worlds possible; for example, the evaluator may be simply introduced to the materials produced, but not to the people producing them, and his comments studied by the director with an eye to feeding back any fundamental and serious criticisms, but withholding the others until some later stage in the curriculum development activities where, for example, an extensive process of revision is about to begin. But these are practical considerations; there remain two more fundamental kinds of objection that should be mentioned briefly, of which the first is central to Atkin's misgivings.

No one who has been involved in the field-testing of a new curriculum has failed to notice the enormous variability in its appeal to students, often unpredictable from their previous academic performance. The child already interested in bird-watching will find one approach to biology far more attractive than another. Similarly, for some children the relevance of the material to problems with which they are familiar will make an enormous difference to their interest, whereas for others the properties of the hexaflexagon or the Moebius strip are immediately fascinating. More fundamentally, the structuring of the classroom situation may wholly alter the motivation for different students in different ways; the non-directive style of treatment currently regarded as desirable, partly for its supposed connection with the inductive approach, is totally unstimulating for some children, although an aggressive, competitive, critical interaction will get them up and running. In the face of this kind of variation, we are often committed to the use of the very blunt evaluation instrument of the performance, on tests, of the class as a whole. Even if we break this down into improvements in individual performances, we still have not fully exploited the potentialities of the material, which would be manifested only if we were to select the right material and the right instructional technique for a child with a particular background, attitudes, interests and abilities. Perhaps, the evaluation skeptic suggests, it is more appropriate to place one's faith in the creative and academically impeccable curriculum maker, using the field tests simply to make sure that it is possible to excite and teach students with the material, under appropriate circumstances. That is, our criterion should be markedly improved performance by some, even by a substantial number, rather than by the class as a whole. To this the evaluator must reply by asking whether one is to disregard possibilities such as serious lack of comprehensibility to students at this age-level, a marked deterioration of performance in some of the students more than offsetting the gains, the possibility that it is the pedagogical skill or enthusiasm of the teacher that is responsible for the success in the field tests and not the materials? The material is to go out to other teachers; it must be determined whether it will be of any use to them. To answer these questions - and indeed for the field tests themselves - a professional job in evaluation is necessary.

We can learn something important from this criticism, however. We must certainly weigh seriously the opinions of the subject matter expert as to the flavor and quality of the curriculum content. Sometimes it will be almost all we have to go on, and sometimes it will even be enough for some decisions. It should in any event be seriously considered and sometimes heavily weighted in the evaluation process, for the absence of supporting professional consensus of this kind is often adequate grounds for complete rejection of the material.

Finally, there is the objection that hovers in the background of many of these discussions, the uneasy feeling that evaluation necessitates making value judgements, and that value judgements are essentially subjective and not scientific. This is about as intelligent a view as the view that statements about oneself are essentially subjective and hence incapable of rational substantiation. Some value judgements are essentially assertions about fundamental personal preferences and as such are factual claims which can be established or refuted by ordinary (though sometimes not easy) procedures of psychological investigation. But the process of establishing them does not show that it is right or wrong to hold these values; it only shows that it is true that somebody does or does not hold them. Another kind of value judgement is the assessment of the merit or comparative merit of some entity in a clearly defined context where this amounts to a claim that its performance is good or better than another's on clearly identifiable and clearly weighted criterion variables. With respect to value judgements of this kind, it is not only possible to find out whether or not they are believed by the individuals who assert them, but it is also possible to determine whether it is right or wrong to believe them. They are simply complex conflations of various performance ratings and the weightings of the various performances; it is in this sense that we can correctly assert that the Bulova Accutron is the best wrist chronometer currently available or that a particular desk dictionary is the best one for somebody with extensive scientific interests. Finally, there are value judgements in which the criteria themselves are debatable, a type of value judgement which is only philosophically the most important of all and whose debatability merely reflects the fact that important issues are not always easy ones. Examples of this would be the assertion that the most important role of evaluation is in the process of curriculum writing, or that the I.Q. test is an unfortunate archaism, or that the Copenhagen interpretation of quantum physics is superior to any alternative. In each of these cases, the disputes turn out to be mainly disputes about what is to count as good, rather than to be arguments about the straightforward 'facts of the situation', i.e. what is in fact good. It is immature to react to this kind of judgement as if it is contaminated with some disgusting disease; the only proper reaction is to examine the reasons that are put forward for them and see if and how the matter may be rationally discussed.

It is sometimes thought that in dealing with people, as we must in the field of education, we are necessarily involved in the field of moral value judgements, and that these really are essentially subjective. But in the first place value judgements about people are by no means necessarily moral, since they may refer to their health, intelligence and achievements; and secondly, even if they are moral, we are all presumably committed to one moral principle (the principle of the equality of rights of men) and by far the greater part of moral discourse takes place within the framework of this assumption, and is simply a rational elaboration of it in combination with complicated judgements about the consequences of alternatives. So, unless one is willing to challenge this axiom, or to provide rational support for an alternative, even moral value judgements are within the realm of rational debate. And even if one does challenge this axiom, a strong case can be made for its rational superiority over any alternatives. But whatever the outcome of such a discussion, the facts that some evaluation is moral evaluation and that some moral evaluation is controversial, do not conjointly imply the least degree of support for the conclusion that curricular evaluation is less than a fully objective activity of applied science.

4. Evaluation versus Process Studies.

In the course of clarifying the concept of evaluation it is important not to simplify it. Although the typical goals of evaluation require judgements of merit and worth, when somebody is asked to evaluate a situation or the impact of certain kinds of materials on the market, then what is being called for is an analytical description of the process, usually with respect to certain possible causal connections. In this sense it is not inappropriate to regard some kinds of process investigation as evaluation. But the range of process research only overlaps with and is neither subsumed by nor equivalent to that of evaluation. We may conveniently distinguish three types of process research, as the term is used by Cronbach and others.

1. The non-inferential study of what actually goes on in the classroom. Perhaps this has the most direct claim to being called a study of the process of teaching (learning etc.). We might for example be interested in the amount of time that the teacher talks, the amount of time that the students spend in homework for a class, the proportion of the dialogue devoted to explaining, defining, opining, etc. (B.O. Smith & Milton Meux). The great problem about work like this is to show that it is worth doing, in any sense. Some pure research is idle research. The Smith and Meux work is specifically mentioned because it is clearly

original and offers promise in a large number of directions. It is difficult to avoid the conclusion, however, that most process research of this kind in education, as in psychotherapy, is fruitful at neither the theoretical nor the applied level.

2. The second kind of process research involves the investigation of causal claims ("dynamic hypotheses") about the process. Here we are interested in such questions as whether an increase of time spent on class discussions of the goals of a curriculum at the expense of time spent on training drills leads to improved comprehension in (a) algebra, (b) geography, etc. This kind of hypothesis is of course a miniature limited-scope 'new instrument' project. Another kind looks for the answer to such questions as, Is the formation of sub-group allegiance and identification with the teacher facilitated by strong emphasis on pupil-teacher dialogue? The identifying feature of this sub-group of process hypotheses is that the dependent variables are either ones which would not figure amongst the set of criteria we would use in a summative evaluation study (though we might think of them as important because of their bearing on improved teaching techniques) or they are only a sub-group of such a set.

Process hypotheses of this second kind are in general about as difficult to substantiate as any 'outcome' hypothesis, i.e. summative evaluation. Indeed they are sometimes harder to substantiate because they may require identifying the effects of only one of several independent variables that are present, and ordinary matching techniques to take care of the others are extremely hard - though usually not impossible - to apply. The advantage of some summative evaluation is that it is concerned with evaluating the effects of a whole teacher-curriculum package and has no need to identify the specific agent responsible for the overall improvement or deterioration. That advantage lapses when we are concerned to identify the variance due to the curriculum as opposed to the teacher.

3. Formative Evaluation. This kind of research can be called process research, but it is of course simply outcome evaluation at an intermediate stage in the development of the teaching instrument. The distinction between this and the first kind of dynamic hypothesis mentioned above is twofold. There is a distinction of role; the role of formative evaluation is to discover deficiencies and successes in the intermediate versions of a new curriculum; the role of dynamic hypothesis investigation is terminal; it is to provide the answer to an important question about the mechanism of teaching. And there is a distinction in the extent to which it matters whether the criteria used are an adequate analysis of the proper goals of the curriculum. The dynamic hypothesis study has no obligation to this; the formative evaluation does. But the two types of study are not always sharply distinct. They both play an important role in good curriculum research.

Now of course it is true that anybody who does an experiment of any kind at all should at some stage evaluate his results. It is even true that the experiment itself will usually be designed in such a way as to incorporate within itself procedures for evaluation of the results - e.g. by using an 'objectively validated' test, which has a certain kind of built-in comparative evaluation in the scoring key. None of this shows that most research is evaluation research. In particular, even process research is not all evaluation research. That interpretation of data can be described as evaluation of results does not show that the interpretations (and the explanations) are about the merit of a teaching instrument. They may be about the temporal distribution of various elements of the instrument etc. Such points are obvious enough, but a good deal of the comment pro and con evaluation research betokens considerable lack of clarity about its boundaries, whose admitted imprecision is really quite slight.

5. Evaluation versus Estimation of Goal Achievement.

One of the reactions to the threat of evaluation, or perhaps to the use of over-crude evaluative procedures, was the extreme relativization of evaluation research. The slogan became "How well does the course achieve its goals?" instead of "How good is the course?". It is of course obvious that if the goals aren't worth achieving then it is uninteresting how well they are achieved. The success of this kind of relativism in the evaluation field rests entirely upon the premise that judgements of goals are value judgements of a non-objective kind. No doubt some of them are; but this in no way indicates that the field is one in which objectivity is impossible. An American History curriculum, K-14, which consisted in the memorisation of names and dates would be absurd - it could not possibly be said to be a good curriculum, no matter how well it attained its goals. Nor could one which led to absolutely no recall of names and dates.

A 'Modern Math' curriculum for general use which produced high school graduates largely incapable of reliable addition and multiplication would be simply a disgrace, no matter what else it conveyed. This kind of value judgement about goals is not beyond debate, but good arguments to the contrary have not been forthcoming so far. These are value judgements with excellent backing. Nor is their defensibility due to their lack of specificity. Much more precise ones can be given just as excellent backing; a physics curriculum which does not discuss the kinetic theory at any stage would be deficient, no matter how well it achieved whatever goals it had. And so on.

Thus evaluation proper must include, as an equal partner with the measuring of performance against goals, procedures for the evaluation of the goals. That is, if it is to have any reference to goals at all. In the next two sections we will discuss procedures of evaluation that involve reference to goals and procedures which short-circuit such reference. First it should be pointed out that there is a complete difference between maintaining that judgement of goals is part of evaluation, i.e. that we cannot just accept anyone's goals, and maintaining that these goals should be the same for every school, for every school district, for every teacher, for every level, etc. It is entirely appropriate that a school with primarily vocational responsibilities should have somewhat different goals from those of a school producing 95% college-bound graduates. It just does not follow from this that the people who give the course or run the school or design the curriculum can be regarded as in any way immune from criticism in setting up their goals. A great deal of the energy behind the current attempts to reform the school curriculum springs straight out of the belief that the goals have been fundamentally wrong, that life-adjustment has been grossly overweighted etc. To swing in the opposite direction is all too easy, and in no way preferable.

The process of relativization, however, has not only led to over-tolerance for over-restrictive goals, it has also led to incompetent evaluation of the extent to which these are achieved. Whatever one's views about evaluation, it is easy enough to demonstrate that there are very few professionally competent evaluators in the country today. The U.S. Office of Education's plans for Research and Development centres, relatively modest in terms of the need, are probably unfulfillable because of the staffing problem, and the heavily financed evaluation projects already in existence are themselves badly understaffed in the evaluation side, even on the most conservative view of its role. Moreover the staff are themselves very well aware of their limitations, and in-service training projects for them are badly needed. The very idea that every school system, or every teacher, can today be regarded as capable of meaningful evaluation of their own performance is as absurd as the view that every psychotherapist today is capable of evaluating his work with his own patients. Trivially, they can learn something very important from carefully studying their own work; indeed they can identify some good and bad features about it. But if they or someone else need to know the answers to the important questions, whether process or outcome, they need skills and resources which are conspicuous by their absence at the national level.

6. Instrumental versus Consequential Evaluation.

Two basically different approaches to the evaluation of a teaching instrument are possible. If you want to evaluate a tool, an instrument of another kind, say an axe, you might study its head design, the arguments for the weight distribution used, the steel alloy in the head, the grade of hickory in the handle, etc., or you might just study the kind and speed of the cuts it makes. (In either case, the evaluation may be either summative or formative, for these are roles of evaluation not procedures for doing evaluation.)

The first approach involves an appraisal of the instrument itself; in the case of a particular course, this would involve evaluation of the content, goals, grading procedures, teacher attitude, etc. We shall call this kind of approach instrumental evaluation. The second approach proceeds via an examination of the effects of the teaching instrument on the pupil, and these alone. It involves an appraisal of the differences between pre- and post-tests, between experimental group tests and control group tests, &c., on a number of criterial parameters. We can call this consequential evaluation.

Referring to the debates between Christians about the foundations of their faith, adherents of the second approach might be inclined to refer to it as the fundamentalist approach, by comparison with the theological approach of the first alternative. Defenders of the second alternative would support this kind of labelling by arguing that all that really counts are the effects of the course on the pupils and appeal to the evaluation of goals and content is defensible only in so far as are evaluations of these really correlates with consequential evaluations. Since these correlations are largely a priori in our present state of knowledge, the fundamentalist argues, the theologian is too much an armchair evaluator. The 'theologian', on the other hand, is likely to counter by talking about values that do not show up in the outcome study to which the fundamentalist restricts himself, and the importance of these in the overall assessment of teaching instruments; he is likely to exemplify this claim by reference to qualities of a curriculum such as elegance, modernity, integrity, etc., which can best be judged by the academic experts in the fields in question.

The possibility arises that an evaluation involving some weighting of instrumental criteria and some of consequential criteria might be a worthwhile compromise. There are certain kinds of evaluation situation where this will be so, but before any assessment of the correct relative weighting is possible it is necessary to look a little further into the difficulties with the two alternatives. In this section we will look at the basic requirements on an instrumental study, in the next examine a currently important disagreement about two types of consequential study, and in the light of our conclusions there we shall be able to say something about the relative merits of instrumental and consequential evaluations.

To recapitulate, it was maintained in the preceeding section that evaluation in terms of goal-achievement is typically a very poor substitute for good summative evaluation. If we are going to evaluate in a way that brings in goals at all, then we shall typically have some obligation to evaluate the goals. As the fundamentalist reminds us, summative evaluation does not necessarily involve any reference to the goals at all, if we do it his way. Indeed one of the charms of the fundamentalist's case is the lack of charm, indeed the messiness, of an adequate instrumentalist design.

A major difficulty with goal-mediated evaluation, which we shall take as the principal example of an instrumentalist approach, lies in the formulation of the goals. In the first place the espoused goals of a curriculum-maker are often not the implicit goals of his curriculum. Moreover, it is not always the case that this kind of error should be corrected in favor of the espoused goals by revising the curriculum, or in favor of the implicit goals by revising the espoused goals. How do we decide which should receive precedence? Even if we were able to decide this, there is the perennial headache of translating the description of the goals that we get from the curriculum-maker or the curriculum-analyst into testable terms. Many a slip occurs between this cup and that lip.

In addition to this, there is the problem already mentioned, that pressure on a writer to formulate his goals, to keep to them, and to express them in testable terms, may enormously alter his product in ways that are certainly not always desirable. Perhaps the best way of handling this third problem is to give prospective curriculum-builders an intensive course in evaluation techniques and problems prior to their commencing work. Such a course would be topic neutral, and would thereby avoid the problems of criticism of one's own 'baby'. Interaction with a professional evaluator can then be postponed substantially and should also be less anxiety-provoking. Short courses of the kind mentioned should surely be available for subsidized attendance every summer at one or two centers in the country. Ignoring any further consideration of the problem of in-group harmony, and this proposal for improving formative evaluation, we can turn to the main difficulty.

6.1 Practical Suggestions for Goal-Mediated Evaluation.

Any curriculum project has some kind of objectives at the very beginning. Even if these are only put in terms of producing a more interesting, or more up-to-date treatment, there has to be some kind of grounds for dissatisfaction with the present curriculum in order to provide a concept of the project as a worthwhile activity. Usually something rather more specific emerges in the course of planning discussions. For example, the idea of a three-track approach, aimed at various kinds of teacher or student interest may emerge out of a rather explicit discussion of the aims of the project, from which it becomes clear that three equally defensible aims can be formulated which will lead to incompatible requirements on the curriculum. The fact that these aims can be seen as incompatible makes clear that they must have fairly substantial content. Another typical content presupposition refers to coverage; it is recognised from the beginning that at least certain topics should be covered, or if they are not then there must be some compensatory coverage of other topics.

At this early stage a member or members of the project team must be appointed to the task of goal-formulation. Many of the objections to this kind of activity stem from reactions to over-rigid requirements on the way in which goals can be formulated at this stage. Any kind of goal on which the group agrees, or even those which they agree should be considered seriously as a possibility in the developing stage, should be listed at this point, but none of them should be regarded as absolute commitments in any way - simply as reminders. It is not possible to overlook the unfortunate examples of projects in which the creative urge has outdistanced reality restraints; it has to be faced from the beginning that too gross a divergence from a certain minimum coverage is going to make the problem of adoption insuperable. If, on the other hand, the risk of negligible adoptions is tolerable, then the goals of the project should be formulated so as to make this clear. Having market-type goals such as substantial adoption on the list is in no way inappropriate: one can hardly reform education with curricula that never reach the classroom.

As the project develops, three types of activities centering around the formulation of goals should be distinguished and encouraged. In the first place the goals as so far formulated should be regularly re-examined and modified in the light of changes in the actual activities, where it is felt that these changes have led to other, more valuable results. Even if no modification seems appropriate, the re-examination will always serve the useful purpose of reminding the writers of overall goals. Secondly, work should be begun on the construction of a test-question pool. Progress tests will be beginning, and the items in these can be thrown into this pool. The construction of this pool is the construction of the operational version of the goals. Consequently it should be scrutinised at the same time as re-examination of goals occurs. Even though the project is only at the stage of finishing the first unit of a projected ten-unit curriculum, it is entirely appropriate to be formulating questions of the kind that it is proposed to include in the final examination on the final unit, or for that matter, in a follow-up quiz. It is a commonplace that in the light of formulating such questions, the conception of the goals of the course will be altered. It is undesirable to require that substantial time be given to this activity, but it is typically not 'undue influence' to encourage thinking about course goals in terms of "What kind of question would tap this learning achievement in the final examination or in a follow-up test?"

At times the answer to this will rightly be "None at all!", for not all values in a course manifest themselves in the final or later examinations. But where they do not thereby manifest themselves, some indication should be given of the time and manner in which they might be expected to be detectable; as in career choices, adult attitudes, etc.

The third activity that should commence at some intermediate stage is that of getting some external judgement as to the cohesiveness of the alleged goals, the actual content, and the test question pool. There is no need at all for the individual judge at this task to be a professional evaluator, and professional evaluators are frequently extremely bad at this. A good logician, an historian of science, a professional in the subject-matter field, an educational psychologist, or a curriculum expert, may be good at this or again they may not. The necessary skill, a very striking one when located, is not co-extensive with any standard professional requirement. This is an area where appointments should not be made without trial periods. It is worth considering whether the activities of this individual, at least in a trial period, may be best conducted without face-to-face confrontation with the project team. A brief written report may be adequate to indicate the extent of possible useful information from the source at this stage. But at some stage, and the earlier the better, this kind of activity is essential if gross divergences between (a) espoused, (b) implicit, and (c) tested-for goals are to be avoided. Not only can a good analyst prevent sidetracking of the project by runaway creative fervor, misconceptions of its actual achievement, etc. but he can provide a valuable stimulus to new lines of development. Ultimately, the justification of psychotherapy does not lie in the fact that the analyst felt he was doing the patient some good, but in the fact that he was; and the same applies to curricular research.

Supposing that this procedure is followed throughout, we will end up with an oversize question pool which should then be examined for comprehensiveness as well as specificity. That is, one should be prepared to say that any significant desired outcome of the course will show up on the answers to these questions; and that what does show up will (normally) only come from the course. Possession of this pool has various important advantages. In the first and second place, it is an operational encapsulation of the goals of the course, if the various cross-checks on its construction have been adequate, which can be used to give the students an idea of what is expected of them as well as to provide a pool from which the final examinations can be constructed. In the third place it can be used by the curriculum-developer to get an extremely detailed picture of his own success (and the success of the cross-checks on pool construction) by administering a different random sample of questions from this pool to each student in a curriculum-check, instead of administering a given random sample to every student as justice requires in a final examination.[1]

What has been described is the bare bones of an adequate mediated evaluation. Now we have made some reference to content characteristics as one of the types of goal, because it is frequently the case that a particular curriculum group argues that one of the merits of its output is its superiority as a representation of contemporary advanced thinking about the subject. The natural way to test this is to have the course read through by some highly qualified experts in the field. It is obvious that special difficulties arise over this procedure. For the most that we can learn from this is that the course does not contain any lies, any distortions of the best contemporary views, or gross deficiencies with respect to them. There remains the question, as the fundamentalist would be the first to point out, of the extent to which the material is being communicated. Even a course with gross oversimplifications, professionally repugnant though it may be to the academic expert, may be getting across a better idea of the truth than its highbrow competitor. The amount of transferred material we infer from the elaborate apparatus of the final test, follow-ups, attitude inventories etc., some details of which are elaborated in a later section. The real advantage of the preceding methodology is to provide a means for making it possible to convert a set of results on the tests into an absolute evaluation, by making reasonably sure that the tests test the goals, one of which may be professional modernity, which may be partly judged by expert reports on the text material, in so far as the tests show this to be transferred fairly uniformly.

---

[1] See Cronbach, ibid. p.242.

A number of further refinements on the above outline are extremely desirable, and in any serious study necessary. Essentially, we need to know about the success of three connected matching problems; first, the match between goals and course content, second, the match between goals and examination content, and third, the match between course content and examination content. Technically we only need to determine two of these in order to be able to evaluate the third; but in fact there are great advantages in attempting to get an estimate of each independently, in order to reduce the error range. We have talked as if one person or group might make each of these matching estimates. It is clearly most desirable that they should all be done independently, and in fact duplicated by independent workers. Only in this way are likely to be able to track down the real source of disappointing results. Even the P.S.S.C. study, which has been as thoroughly tested as most recent curriculum projects, has nowhere approached the desirable level of analysis indicated here.

In general, of course, the most difficult problem in tests and measurement theory is the problem of construct validity, and the present problem is essentially an exercise in construct validity. The problem can be ignored, but only by someone who is prepared to accept immediately the consequence that their supposed goals cannot be regarded as met by the course, or that their examinations do not test what the course teaches, or that the examinations do not test the values/materials that are supposed to be imparted by the course. There are, in practice, many ways in which one can implement the need for comparisons here described; the use of Q-sorts and R-sorts, matching and projective tests for the analysts etc. In one way or another the job has to be done - if we are going to do a mediated evaluation.

6.2 The Possibility of Bypassing Goal Evaluation.

The pure consequentialist, the 'fundamentalist', tends to watch the intricacies of this kind of experimental design with glee, for he believes that the whole idea of bringing in goal- or content-assessment is not only an irrelevant but an extremely unreliable procedure for doing the job of course evaluation. In his view it isn't very important to examine what a teacher says he is doing, or what the students say he is doing (or they are learning), or even what the teacher says in class; the only important data is what the student says (does, believes, etc.) at the end of the course that he wouldn't have said at the beginning (or, to be more precise, would not have said at the end if he had not taken this course). In short, says the fundamentalist, let's see what the course does, and let's not bother with the question of whether it had good intentions.

But the fundamentalist has difficulties of his own. He cannot avoid the construct validity issue entirely, that is, he cannot avoid the enormous difficulties involved in correctly describing at a useful level of generality what the student has learned. It is easy enough to give the exact results of the testing in terms of the percentage of the students who gave certain answers to each specific question: but what we need to know is whether we can say, in the light of their answers, that they have a _better_ understanding of the elements of astronomy, or the chemical-bond approach to chemistry, or the ecological approach to biology. And it is a long way from data about answers to questions, to that kind of conclusion. It is not _necessary_ for the route to lie through a discussion of goals - the fundamentalist is quite right about this. But if it does _not_ lie through a discussion of goals, then we shall not have available the data that we need (a) to distinguish between importantly different explanations of success or failure, (b) to give reasons for using the new text or curriculum to those whose explicit aim _is_ the provision of better understanding of the chemical-bond approach. For example, if we attempt a fundamentalist approach to evaluating a curriculum, and discover that the material retained and regurgitated by the student is regarded as grossly inadequate by the subject-matter specialists, we have no idea whether this is due to an inadequacy in the goals of the curriculum-makers, or to imperfections in their curriculum with respect to these goals, or to deficiencies in their examinations with respect to either of the preceding. And thus we cannot institute a remedial program - our only recourse is to start all over. Fundamentalism can be a costly simplification.

Suppose that we follow a fundamentalist approach and have the students' performance at the end of the course, and only this, rated by an external judge. Who do we pick for a judge? The answer to that question will apparently reveal a commitment on our own part to certain goals. The evaluator will have to relate the students' performance to _some_ criterion, whether it is his conception of an adequate _professional_ comprehension, or what he thinks it is reasonable to expect a _tenth-grader_ to understand, or what somebody should understand who will not continue to college etc. The fundamentalist is right in saying that we can dispense with any discussion of goals and still discover exactly what students have learnt, and right to believe that the latter is the most important variable; but he is mistaken if he supposes that we can in general give the kind of description of what is learnt that is valuable for our purposes without any reference to goals. At some stage, someone is going to have to decide what counts as adequate comprehension for students at a particular level, for a particular subject, and then apply this decision to the non-evaluative descriptions of what the students have learnt, in order to come up with the overall evaluation. At this stage of the debate between the supporter of fundamental and mediated evaluation, the latter would seem to be having the best of it, particularly since there are certain goals that can be (a) incorporated into a course (b) judged as worth incorporating by subject-matter authorities, but which (c) are not such as to show up in an appropriate kind of final examination at the end of a particular year. But the issue is not so one-sided; the fundamentalist is performing an invaluable service in reminding us of the potential irresponsibility of producing "elegant", "up-to-date", "rigorous" curricula if these qualities are not coming through to the students. We can take them on faith insofar as they are recognised as being the frosting on the cake; but we can't take the food-value of the cake on faith. The amount of goal analysis that is absolutely necessary in order to provide a summative evaluator with the basis for a value-judgement about the curriculum is very, very little compared with the amount that a thorough mediated evaluation involves. It is, after all, more important to put time and money into deciding whether what the student has acquired is a misconception of the nature of electric current than whether the curriculum-writer has inadvertently incorporated some minor misconception of it into his curriculum. The real alternative which the fundamentalist presents is the use of an academic evaluator who is asked to look at the exact performance of the class on each question and at the pool from which the questions were drawn, and from these directly assess the adequacy of the course to the subject as he sees it. Such an evaluator makes his evaluations by reference to a criterion of merit, but this is not the same as saying that he presupposes something about the goals of the course. He may think it unlikely that a course should be much good (in terms of his criteria) unless it had his criteria as explicit or implicit goals, but he is not at all committed to such a claim. He is committed to the view that certain goals are or would be desirable, but they may be goals that no course-maker has ever employed. So there is no contradiction in the fundamentalist view that we do not have to have or evaluate goals in order to evaluate a course, and he is certainly right in believing that bringing them in makes for an invalid or very complex design. Yet sometimes we have good practical reasons for doing so.

In conclusion, it should be clear that a strong case can be made for incorporating the procedure described above as part of any good curriculum project, whether or not we use mediated evaluation. Doing so will of course help to make a good mediated evaluation feasible. In addition, however, it should be noted that an equally thorough analysis is required of the results of the students' tests, and not only of the course content. It is not at all adequate to go to great trouble setting up and cross-analyzing the goals, tests, and content of a curriculum and then attempt to use a percentage figure as the indication of goal achievement (unless the figure happens to be pretty close to 100% or 0%). This kind of gross approach is no longer acceptable as evaluation. The performance of the students on the final tests, as upon the tests at intermediate stages, must be analysed in order to determine the exact locations of shortcomings of comprehension, shortages of essential facts, lack of practice in basic skills etc. Percentages are not very important. It is the nature of the mistakes that is important in evaluating the curriculum, and in rewriting it. The technique of the large question pool provides us with an extremely refined instrument for locating deficiencies in the curriculum. But this instrument can only be exploited fully if evaluation of the results is itself handled in a refined way, with the same use of independent judges, putative generalizations about the nature of the mistakes being cross-matched etc. It should be clear that the task of proper evaluation of curriculum materials is an enormous one. The use of essay type questions, the development and use of novel instruments, the use of reports by laboratory-work supervisors, the colligation of all this material into specially developed rating schemata, all of this is expensive and time-consuming. In a later section some consideration of the consequences of this picture of the scale of evaluation activities will be undertaken. At this point, however, it becomes necessary to look into a further and final divergence of approaches.

7. Comparative versus Non-Comparative Evaluation.

The history of attempts to evaluate recent curricular reforms has been remarkably uniform; comparing students taking the old curriculum with students taking the new one, it usually appears that students using the new curriculum do rather better on the examinations designed for that curriculum and rather worse on those designed for the old curriculum, while students using the old curriculum perform in the opposite way. Certainly, there is a remarkable absence of striking improvements on the same criteria (with some exceptions, of which the most notable is the performance of students in studies of good programmed texts). Initially, one's tendency is to feel that the mountain has laboured and brought forth a mouse - and that it is a positive mouse and not a negative one entirely depends upon the evaluation of the goals (and hence of the examinations). A legitimate reaction is to look very seriously into the question of whether one should not weight judgement of content and goals by subject-matter experts as being a great deal more important than small differences in level of performance on these criteria. If we do this, then relatively minor improvements in performance, on the right goals, become very valuable, and in these terms the new curriculum looks considerably better. Whether this alteration of weights can really be justified is a matter that needs very serious investigation; it requires a rather careful analysis of the real importance to the understanding and use of contemporary physics, as it is seen by physicists, of the missing elements in the old curriculum. It is all too tempting to feel that the re-weighting must be correct because one is so thoroughly convinced that the new course is better.

Another legitimate reaction is to wonder whether the examinations are really doing a good job testing the depth of understanding of the people trained on the new curriculum. Here the use of the over-size question pool becomes extremely important. Cronbach speaks of a 700 item pool (without flinching!) and this is the kind of order of magnitude that makes sense in terms of an exhaustive evaluation of a one or two-year curriculum. Whether this reaction reveals a legitimate basis for increasing the measure of importance of the difference between the students groups using the new and old curricula will depend upon the results of further tests using a thoroughly justified and much enlarged pool. Again, it is going to be tempting to put items into the pool that reflect mere differences of terminology in the new course, for example. Of course if the pool consists mainly of questions of that kind, the new curriculum-students will do much better. But their superiority will be entirely illusory. Cronbach warns us against this risk of course-dependent terminology, although he goes too far in segregating understanding from terminology (this point is taken up below). So here, too, we must be certain to use external evaluators in the construction or assessment of the question pool.

Other illegitimate reactions run from the charming suggestion that such results simply demonstrate the weaknesses of evaluation techniques, to a more interesting suggestion implicit in Cronbach's paper. He says:

> "Since group comparisons give equivocal results, I believe that a formal study should be designed primarily to determine the post-course performance of a well-described group, with respect to many important objectives and side-effects." [1]

Notice that Cronbach is not producing an alternative to mediated evaluation, in the way that the fundamentalist is; Cronbach explicitly includes reference to pre-evaluated objectives i.e. important objectives. He is apparently about to suggest a way in which we can avoid comparison, not with goals or objectives, but with another group, supposedly matched on relevant variables. What is this non-comparative alternative procedure for evaluation? He continues;

> "Ours is a problem like that of the engineer examining a new automobile. He can set himself the task of defining its performance characteristics and its dependability. It would be merely distracting to put his question in the form: 'Is this car better or worse than the competing brand?'"

It is perfectly true that the automobile engineer might just be interested in the question of the performance and dependability of the new automobile.

---

[1] This and the succeeding quotation are from p.238.

But no automobile engineer ever has had this pure interest, and no automobile engineer ever will have it. Objectives do not become "important" except in a practical context. Unrealistic objectives are not important. The very measures of the performance and dependability of an automobile and our interest in them spring entirely from knowledge of what has and has not so far proved possible, or possible within a certain price-class, or possible with certain interior space, or with a certain overall weight etc. The same applies in the field of curriculum development. We already have curricula aimed at almost every subject known to man, and there isn't any real interest in producing curricula for curricula's sake; to the extent that there is, there isn't any interest in evaluating them. We are interested in curricula because they may prove to be better than what we now have, in some important way. We may assign someone the task of rating a curriculum on certain variables, without asking them simultaneously to look up the performance of other curricula on these variables. But when we come to evaluate the curriculum, as opposed to merely describing its performance, then we inevitably confront the question of its superiority or inferiority to the competition. To say it's a valuable contribution, a desirable or useful course, even to say - in the usual context - that it's very good, is to imply relative merit. Indeed the very scales we use to measure its performance are often percentile scales or others with a built-in comparison.

There are even important reasons for putting the question in its comparative form immediately. Comparative evaluations are often very much easier than non-comparative evaluations, because we can often use tests which yield differences instead of having to find an absolute scale and then eventually compare the absolute scores. If we are discussing chess-teaching courses, for example, we might match two groups for background variables, and then let them play eachother off in a round-robin tournament. Attempting to devise a measure of skill of an absolute kind would be a nightmare, but we might easily get consistent and significant differences from this kind of comparative evaluation. Cronbach is not making the fundamentalist's mistake of thinking that one can avoid reference to goals; but he is proposing a kind of neo-fundamentalism which underestimates the implicit comparative element in any field of social engineering including automobile assessment and curriculum evaluation.

Cronbach continues in this paragraph with a line of thought about which there can be no disagreement at all; he points out that in any cases of comparisons between importantly different teaching instruments, no real understanding is gained from the discovery that one of them is notably superior to the other: "No one knows which of the ingredients is responsible for the advantages". But understanding is not our only goal in evaluation. We are also interested in questions of support, encouragement, adoption, reward, refinement etc. And these extremely important questions can be given a useful though in some cases not a complete answer by the mere discovery of superiority. It will be recalled that in an earlier section we argued that the fundamentalist position suffers by comparison with the supporter of mediated evaluation in that his results will not include the data we need in order to locate sources of difficulty etc. Here Cronbach is arguing that his non-comparative approach will be more likely to give us the data we need for future improvement. But this is not in any way an advantage of the non-comparative method as such. It is simply an advantage of methods in which more variables are examined in more detail. If we want to pin down the exact reasons for differences between programs, it is quite true that "small-scale, well-controlled studies can profitably be used to compare alternative versions of the same course" whereas the large-scale overall comparison will not be so valuable.

But that in no way bears on the question whether we have any alternative to comparative studies at some point in our evaluation procedures. In short this is simply an argument that one needs more control groups, and possibly more short-run studies in order to get explanations, than one needs for overall evaluation. It is incontestible; but it does not show that for the purposes of overall evaluation we can or should avoid overall comparison.

One might put the point in terms of the following analogy; in the history of automobile engine design there have been a number of occasions when a designer has turned out an engine that was quite inexplicably superior to the competition - the Kettering GM V8, the Coventry Climax and the Weslake Ford Conversions are well-known examples. At least thirty variables are involved in the design of any new engine and for a long time after these had been in production nobody, including the designer, knew which of them had been mainly responsible for the improvement. But the decision to go into production, the decision to put the further research into the engine that led to finding out what made it great, indeed the beginning of a new era in engine design, required only the comparative evaluation. You set a great team to work and you hope they are going to strike gold; after that you stake your claim and start trying to work out the configuration of the lode. This is the way we have to work in any field where there are too many variables and too little time.

7.1 Practical Procedures in Control-Group Evaluation.

It is a major theme of Cronbach's that control group comparisons in the curriculum game are not really very suitable. We have just seen how his attempt to provide a positive alternative does not develop into a realistic answer in the context of typical evaluation enquiries. It is now appropriate for us to attempt to meet some of the objections that he raises to the control group method if we are to recommend that this be left in possession of the field.

The suggestion that gross comparisons yield only small differences must be met, as indicated above (and as he recommends elsewhere), by increasing the power of the microscope - that is, by increasing the number of items that are being tested, increasing the size of the group in order to get more reliability into differences that do appear, and developing new and more appropriate tests where they seem to be the weakness. But once all this has been said, the fact remains that it is probably the case that we shall have to proceed in terms of rather small differences; that producing large differences will probably require a multiple-push approach, attacking not only the curriculum but the student-grouping procedures, the teacher presentation, the classroom time allocation, and above all the long-term effects that an attack on every subject in the school curriculum will eventually produce for us, a general increase in the level of interest and preparedness. This is not too depressing a prospect, and it is exactly paralleled in that other field in which we attempt to change human behaviour by applying pressure on the subjects for a few hours a week over a period of one or several years - the field of psychotherapy. We are perhaps too used to the discovery of miracle drugs or technological breakthroughs in the aero-space field to realise how atypical this is of progress in general. In the automobile engineering field, to stay with Cronbach's example, it is well known that developing a good established design yields better results than introducing a radical and promising new design in about twice as many cases as engineers under forty are willing to believe. What one may reasonably expect in the way of progress is not great leaps and bounds, but steady improvement. Cronbach says that "formally designed experiments pitting one course against another are rarely definitive enough to justify their cost" but this is just the kind of knowledge that we need to have. If we have really satisfied ourselves that we are using good tests of every criterion variable that matters (and of course we usually have a number in the follow-up series that make this kind of conclusion impossible for a few years) then to discover parity of performance is to have discovered something extremely informative.

Of course, we cannot conclude from this that all the techniques involved in the new curriculum are worthless improvements. We must go on to make the micro-studies that will enable us to see whether any one of them is worthwhile. But we have discovered something very significant. Doing the gross comparative study is going to cost the same whatever kind of results we get, and we have to do it. The real question is whether we stop after discovering an insignificant difference, or continue in the direction of further analytical research, as Cronbach enthusiastically recommends (or incorporate the refinements in the original design which will give us the further answer). The impact of his article is to suggest the unimportance of the control group study, whereas the case can only be made for its inadequacy as a total approach to the whole of curriculum research.[1] We shall here try to provide some practical suggestions for experimental designs that will yield more than a gross comparative evaluation.

A significant part of the reason for Cronbach's despair over comparative studies lies in his recognition that we are unable to arrange for double-blind conditions. "In an educational experiment it is difficult to keep people unaware that they are an experimental group. And it is quite impossible to neutralise the biases of the teacher as those of the doctor are neutralised in the double-blind design. It is thus never certain whether any observed advantage is attributable to the educational innovation as such, or to the greater energy that teachers and students put forth when a method is fresh and 'experimental'." (p.237) But Cronbach despairs too quickly. The analogy in the medical field is not with drug studies, where we are fortunate enough to be able to achieve double-blind conditions, but with psychotherapy studies where the therapist is obviously endowed with enthusiasm for his treatment, and the patient cannot be kept in ignorance of whether he is getting some kind of treatment. If Cronbach's reasoning is correct, it would not be possible to design an adequate psychotherapy outcome study. But it is possible to design such a study, and the way to do it - as far as this point goes - is to make comparisons between a number of therapy groups, in each of which the therapist is enthusiastic, but in each of which the method of therapy is radically different.[1] As far as possible, one should employ forms of therapy in which directly incompatible procedures are adopted. There are already a number on the market which meet this condition in several dimensions, and it is easy enough to develop pseudo-therapies which would be promising enough to be enthusiasm-generating for some practitioners (e.g. newly graduated internists inducted into the experimental program for a short period). The method of differences plus the method of concomitant variations will then enable us to draw straight-forward conclusions about whether enthusiasm is the (or a) major factor in therapeutic success, even though double-blind conditions are unobtainable. Nor is this the only kind of design which can do this; many other devices are available, and ingenious experimenters will doubtless think of still more, to enable us to handle this kind of research problem. There is nothing indispensable about the double-blind study.

---

[1] Yet he does agree with the necessity for making the practical decisions between textbooks and similar instructional materials (p.232), for which nothing less than a valid comparative study is adequate.

---

[1] Other difficulties are discussed in more detail in "The Experimental Investigation of Psychoanalysis" in Psychoanalysis, Scientific Method and Philosophy ed. S.Hook, NYU Press 1959.

Now the curriculum field is even more difficult than the psychotherapy field, because, although the average intelligent patient will accept almost any nonsense as a form of therapy, thanks to the witchdoctor tradition, need to be healed etc. it is not equally easy to convince students and teachers that they are receiving and giving instruction in geometry unless what is going on really is a kind of geometry that makes some sense. And if it is, then interpretation of one of the possible outcomes is ambiguous, i.e. if the two groups do about as well, it may be because enthusiasm does the trick, or because the content is about equally valuable. However, comparative evaluation is still well worthwhile, because if we find a very marked difference between the groups, and are able to arrange for enthusiasm on the part of the teachers and students in both cases, we may be reasonably sure that the difference is due to the curriculum content.

Now it is not particularly difficult to arrange for the enthusiasm matching. Corresponding to the cut-rate therapy comparison group, where the therapy procedures are brainstormed up in a day or two of wild free-associating by the experimenters assisted by a lot of beer and some guilt-ridden eclectic therapists, we set up some cut-rate new curricula in the following way. First, we get two bright graduate students or instructors in (let us suppose) economics, give them a vocabulary list for the tenth grade and pay them $500 a chapter for a translation of Samuelson's text into tenth grade language, encouraging them to use their originality in introducing the new ideas. They could probably handle the whole text in a summer and so for a few thousand dollars, including costs of reproducing pilot materials, we have something we could set up against one of the fancier economics curriculum, based on a great deal of high-priced help and laborious field-testing. Then we find a couple of really bright college juniors, majoring in economics, from different colleges, and give them a summer to turn their recent experience at the receiving end of introductory economics courses, and their current direct acquaintance with the problems of concept grasping in the field, into a curriculum outline, not centered around any particular text, filled in as much as possible, of a brief introduction to economics for the tenth-grade. And for a third comparison group we locate some enthusiasts for one of the current secondary school texts in 'economics' and have them work on a revision of it with the author(s) and in the light of some sampling of their colleagues reactions to the text in class use.

Preferably using the curriculum-makers as teachers (pace State Departments of Education) we then turn them loose on matched comparison groups, in school systems geographically well removed from the ones where we are running the tests on the high-priced spread. We might toss in a little incentive payment in the way of a pre-announced bonus for these groups if they don't get significantly out-scored by the super-curriculum. Now then, if we still get a big difference in favor of the super-curriculum, we have good reason for thinking that we have taken care of the enthusiasm variable. Moreover we don't have to pull this stunt with every kind of subject matter, since enthusiasm is presumably reasonably (though definitely not entirely) constant in its effects across subject matter. At any rate, a modest sampling should suffice to check this.

One of the nice things about this kind of comparative study is that even if we get the ambiguous negligible-difference result, which will leave us in doubt as to whether a common enthusiasm is responsible for the result, or whether a roughly comparable job in teaching economics is being done by all the curricula, we get a nice economic bonus. If we can whomp up new curricula on a shoestring which are going to produce pretty good results, so much the better: we can do it often and thereby keep up the supply of enthusiasm-stoked project directors, and increase the chances of hitting on some really new big-jackpot approach from a Newton of curriculum reform.

Moreover, still on a shoestring, we can settle the question of enthusiasm fairly quickly even in the event of a tie between the various curricula, by dumping them into the lap of some antagonistic and some neutral teachers to use during the next school term, while on the other hand arranging for the original curriculum-makers to lovingly train a small group of highly selected and innovation-inclined teachers to do the same job. Comparisons between the performance of these two new groups and that of the old ones should enable us to pin down the role of enthusiasm rather precisely, and in addition the no-doubt variable immunity of the various curricula to lack of enthusiasm.

A few obvious elaborations of the above procedures, including an opportunity for the novice curriculum-makers to spend a couple of afternoons on field-testing early sections of their new curriculum, to give them some 'feel' for the speed at which students at this level can grasp new concepts, the use of some care in selecting teachers for their conservatism or lethargy, using self-ratings plus peer-ratings plus attitude inventories, would immediately suggest themselves in the case of an actual study.

The enthusiasm 'difficulty' here is simply an example of what we might call disturbance effects, of which the placebo effect in medicine and the Hawthorne effect in industrial and social psychology are well-known instances. In each case we are interested in finding out the effects of a certain factor, but we cannot introduce the factor into the experimental situation without producing a disturbance which may itself be responsible for the observed changes. In the drug field, the disturbance consists in the act of giving the patient something which he considers to be a drug, something which does not ordinarily happen to him, and consequently may produce effects of its own, quite apart from the effects of the drug. In the Hawthorne effect, the disturbance is the disruption of e.g. conditions of work which may suggest to the worker that he is the subject of special study and interest, and this may lead to improved output, not the physical changes in the environment that are the intended parameters under study.

The cases so far mentioned are all ones where the beliefs of the subjects are the mediating factor between the disturbance and the ambiguous effects. This is characteristic in the field of psychology, but - as the term 'disturbance effect' indicates - the situation is not essentially different from that occurring in technological research where we face problems such as the absorption of heat by a thermometer which thereby alters the temperature that it is supposedly measuring. That is, some of the effect observed (which is here the eventual length of the mercury column) is due to the fact that in order to get the effect at all you have to introduce another physical object into proximity with the measured object, the instrument itself having a certain heat capacity, a factor in whose influence you are not interested though in order to find out what you do need to know you eventually have to make an estimate of the magnitude of the disturbance effect. The ingenious double-blind design is only appropriate in certain circumstances, and is only one of many ways in which we can compensate for disturbance effects. It therefore seems unduly pessimistic of Cronbach to suppose that the impossibility of a double-blind in curriculum work is fatal to comparative evaluation. Indeed, when he comes to discuss follow-up studies, he agrees that comparative work is essential (p.240). The conclusion seems obligatory that comparative evaluation, whether mediated or fundamental is the method of choice for evaluation problems.

8. Comparative Evaluation - The Criteria of Educational Achievement.

We may now turn to the problem of specifying in more detail the criteria which should be used in evaluating a teaching instrument. We may retain Bloom's[1] convenient trichotomy of cognitive, affective and motor variables, though we shall often refer to the last two as motivational and physical or non-mental variables, but under the first two of these we shall propose a rather different structure, especially under the knowledge and understanding subdivisions of the cognitive field. It should be stressed at the beginning that the word "knowledge" can be used to cover understanding (or comprehension) and even affective conditions, but that it is here used in the sense in which it can be contrasted with comprehension and experience or valuation, i.e. in the sense in which we think of it as 'mere knowledge'. Comprehension or understanding, by contrast, refers to a psychological state involving knowledge, not of one item, nor of several separate items, but of a field. A field or structure is a set of items related in a systematic way, knowledge of the field involving knowledge not only of the items but of their relations. A field is often open-ended in the sense of having potential reference or applicability to an indefinite number of future examples. In this latter case, comprehension involves the capacity to apply to these novel cases the appropriate rule, rubric or concept. A field may be a field of abstract or practical knowledge, of thought or of skills.

With respect to any field of knowledge we can distinguish between a relatively abstract or conceptual description of the parameters (which are to occupy the role of dependent variables in our study) and a manifestation description, the latter being the next stage towards the specification of the particular tests to be used, which we may call the operational description. It is appropriate to describe the criteria at all three levels, although we finally apply only the third, just as it is appropriate to give the steps of a difficult proof in mathematics, because it shows us the reasons for adopting the particular final step proposed.

---

[1] Taxonomy of Educational Objectives B.S. Bloom editor and others, Vols. I, II, and III (forthcoming).

I have followed the usual practice here in listing positive goals (with the possible exception of the example in 5) but a word of caution is in order. Although most negatively desired effects are the absence of positively desired effects, this is not always true, and more generally it is often true that one may wish to alter the weighting of a variable when it drops below a certain level. For example, we may not be worried if we get no change on socialization with a course that is working well in the cognitive domain, and we may give small credit for large gains in this dimension. But if it produces a marked rise in sociopathic behaviour we may regard this as fatal. Similarly with respect to forgetting or rejection of material in other subject areas etc. Another example is discussed below.

A word about originality; this may be manifested in a problem-solving skill, an artistic skill (which combines motor and perceptual and perhaps verbal skills) and in many other ways. It does not seem desirable to make it a separate criterion.

In general, I have tried to reduce the acknowledged overlap amongst the factors identified in Bloom's analysis, and am prepared to pay a price for this desideratum, if such a price must be paid. There are many reasons for avoiding overlap, of which one of the more important and perhaps less obvious ones is that when the comparative weighting of criteria is undertaken for a given subject, independence greatly simplifies the process, since a straight weighting by merit will overweight the hidden loading factors.

There is still a tendency in the literature to regard factual recall and knowledge of terminology with disdain. But for many subjects, a very substantial score on that dimension is an absolutely necessary condition for adequate performance. This is not the same as saying that a sufficiently high score on that scale will compensate for lack of understanding, even where we use a single index compounded from the weighted scores. There are other subjects, especially mathematics and physics, where knowing the terminology requires and hence guarantees a very deep understanding and terminology-free tests are just bad tests. (cf Cronbach p.245)

8.1 Conceptual Description of Educational Objectives.

1. Knowledge, of
    - A. Items of specific information including definitions of terms in the field.
    - B. Sequences or patterns of items of information including rules, procedures or classifications for handling or evaluating items of information (we are here talking about mere knowledge of the rule and not the capacity to apply it).
2. Comprehension or Understanding, of
    - A. Internal relationships in the field,[1] i.e. the way in which some of the knowledge claims are consequences of others and imply yet others, the way in which the terminology applies within the field; in short what might be called understanding of the intrafield syntax of the field or sub-field.
    - B. Inter-field relations, i.e. relations between the knowledge

---

[1] Typically, 'the field' should be construed more widely than 'the subject' since we are very interested in transfer from one subject to related ones and rate a course better to the extent it facilitates this. In rating applications, we can range very far e.g. from a course on psychology to reactions to commercials showing white-coated men.

claims in this field and those in other fields; what we might call the interfield syntax.

C. Application of the field or the rules, procedures and concepts of the field to appropriate examples, where the field is one that has such applications; this might be called the semantics of the field.

3. Motivation. (Attitude/values/affect)

A. Attitudes towards the course, e.g. acoustics.

B. Attitudes towards the subject, e.g. physics.

C. Attitudes towards the field, e.g. science.

D. Attitudes towards material to which the field is relevant, e.g. increased skepticism about usual advertising claims about 'high fidelity' from miniature radios (connection with 2C above).

E. Attitudes towards learning, reading, discussing it, enquiring in general etc.

F. Attitudes towards the school.

G. Attitudes towards teaching as a career, teacher status etc.

H. Attitudes towards (feelings about etc.) the teacher as a person.

I. Attitude towards class-mates, attitude towards society (obvious further sub-headings).

J. Attitude towards self, e.g. increase of realistic self-appraisal (which also involves cognitive domain).

4. Non-Mental Abilities.

A. Perceptual.

B. Psycho-motor.

C. Motor, including e.g. some sculpting skills.

D. Social skills.

5. Non-Educational Variables.

There are a number of non-educational goals, usually implicit, which are served by many courses and even new courses, and some of them are even justifiable in special circumstances as e.g. in a prison. The crudest example is the 'keeps 'em out of mischief' view of schooling. It is realistic to remember that these criteria may be quite important to parents and teachers even if not to children.

8.2 Manifestation Dimensions of Criterial Variables.

1. Knowledge (in the sense described above) is evinced by

   A. Recital skills.

   B. Discrimination skills.

   C. Completion skills.

   D. Labelling skills.

   Note: Where actual performance changes are not discernible, there may still be some subliminal capacity, manifesting itself in a reduction in re-learning or in future learning to criterion.

2. Comprehension is manifested on some of the above types of performance and also on

   A. Analysing skills, including laboratory analysis skills, other than motor, as well as the verbal analytic skills, exhibited in criticism, précis, etc.

   B. Synthesising skills.

   C. Evaluation skills.

   D. Problem-solving skills (speed-dependent and speed-independent).

3. Attitude manifestations usually involve simultaneous demonstration of some cognitive acquisition. The kinds of instrument involved

are questionnaires, projective tests, Q-sorts, experimental choice situations, and normal lifetime choice situations (choice of college major, career, spouse, friends, etc.). Each of the attitudes mentioned is characteristically identifiable on a passive to active dimension (related to the distinctions expounded on in Bloom, but disregarding extent of systematisation of value system which can be treated under meta-cognitive skills).

4. The Non-Mental Abilities are all exhibited in performances of various kinds, which again can be either artificially elicited or extracted from life-history. A typical example is the capacity to speak in an organised way in front of an audience, to criticise a point of view not previously heard in an effective way etc. (this again connects with the ability conceptually described under 2C).

8.3 Follow-Up.

The time dimension is a crucial element in the analysis of performance and one that deserves an extensive independent investigation. Retention, recall, depth of understanding, extent of imprinting, can all be tested by re-applications of the tests or observations used to determine the instantaneous peak performance, on the dimensions indicated above. However, some follow-up criteria are not repetitions of earlier tests or observations; eventual choice of career, longevity of marriage, extent of adult social service, career success, are relevant and important variables which require case history investigation. But changes of habits and character are often not separate variables, being simply long-term changes on cognitive and affective scales.

8.4 Secondary Effects.

A serious deficiency of previous studies of new curricula has been a failure to adequately sample the teacher population. When perfecting a teaching instrument, we cannot justify generalising from pilot studies unless not only the students but the teachers are fair samples of the intended population. This is one reason for the importance of the studies of interference effects. Just as generalising has been based upon inadequate analysis of the teacher sample, so criterion discussions have not paid sufficient attention to teacher benefits. It is quite wrong to evaluate a teaching instrument without consideration of the effects on the operator as well as on the subjects. In an obvious sense, the operator is one of the subjects.

We may divide secondary effects (i.e. those on others than the students taking the course) into two categories. Direct secondary effects are those arising from direct exposure to the material, and only the teachers and teachers' helpers can be affected in this way. Indirect secondary effects are those effects mediated by someone who exhibits the primary effects.

8.41 Effects on the Teacher.

A new curriculum may have very desirable effects on up-dating a teacher's knowledge, with subsequent pay-off in various ways including the better education of other classes at a later stage, in which he/she may be using either the old curriculum or the new one. Similarly, it may have very bad effects on the teacher, perhaps through induction of fatigue, or failing to leave her any feeling of status or significant role in the classroom etc.

It is easy to itemise a number of such considerations, and we really need a minor study of the taxonomy of these secondary effects under each of their several headings. In particular, what I have called the interference effects e.g. those due to enthusiasm, can be directly valued, as I think they should be - if we include secondary effects in the criteria. Very often the introduction of new curriculum material is tied to teacher in-service training institutes or special in-service training interviews. These of course have effects on the teacher herself with respect to status, self-concept, pay, interests etc, and indirectly on later students. Many of these effects on the teacher show up in her other activities; at the college level there will normally be some serious reduction of research time resulting from association with an experimental curriculum, and this may have results for promotion expectations in either the positive or the negative direction, depending upon departmental policy. All of these results are effects of the new curriculum, at least for a long time, and in certain circumstances they may be sufficiently important to count rather heavily against other advantages. Involvement with curricula of a highly controversial kind may have such strongly damaging secondary effects as to raise questions as to whether it is proper to refer to it as a good curriculum for schools in the social context in which these secondary effects are so bad.

8.42 Indirect Effects on Teacher's Colleagues.

Indirect secondary effects are the effects on people other than those directly exposed to the curriculum: once again they may be highly significant. A simple example of an indirect secondary effect involves other members of the staff who may be called upon to teach less attractive courses, or more courses, or whose load may be reduced for reasons of parity, or who may be stimulated by discussions with the experimental group teachers, etc. In many cases, effects of this kind will vary widely from situation to situation, and such effects may then be less appropriately thought of as effects of the curriculum (although even the primary effects of this, i.e. the effects on the students will vary widely geographically and temporally) but there will sometimes be constancies in these effects which will require recognition as characteristic effects of this particular teaching instrument. This will of course be noticeable in the case of controversial experimental courses, but it will also be significant where the course bears on problems of school administration, relation of the subject to other subjects, and so on. Good evaluation requires some attempt to identify effects of this kind.

8.43 Indirect Effects on Other Students.

Another indirect secondary effect, only partly covered in the effect of the curriculum on the teacher, is the effect on other students. Just as a teacher may be improved by exposure to a new curriculum, and this improvement may show up in benefits for students that she has in other classes, or at a later period using the old curriculum etc., so there may be an effect of the curriculum on students not in the experimental class through the intermediary of students who are. Probably more pronounced in a boarding school, the communication between students is still a powerful enough instrument in ordinary circumstances for this to be a significant influence. The students may of course be influenced in other ways; there may be additions to the library as a result of the funds available for the new course that represent values for the other students etc. All of these are educationally significant effects of the course adoption.

8.44 Effects on Administrators.

The college administrators may be affected by new teaching instruments in various ways; their powers of appointment may be curtailed, if the teaching instrument's efficiency will reduce faculty, they may acquire increased prestige (or nuisance) through the use of the school as an experimental laboratory, they may find this leads to more (or less) trouble with the parents, the pay-off through more national scholarships may be a value to them, either intrinsically or incidentally to some other end, etc. Again, it is obvious that in certain special cases this variable will be a very important part of the total set that are affected by the new instrument, and evaluation must include some recognition of this possibility. It is not so much the factors common to the use of novel material, but the course-specific effects that particularly require estimation and almost every new science or social studies course has such effects.

8.45 Effects on Parents.

Effects on the parents are of course well known, but they tend to be regarded as nuisance-generating effects. On the contrary, many such effects should be regarded as part of the adult education program in which this country is remarkably lacking. In some subjects, e.g. Russian, this is unlikely to have a very significant effect, but in the field of problems of democracy, elementary accounting, and literature, this may be a most important effect.

8.46 Effects on the School or College.

Many of these are covered above, particularly under the heading of effects on the administrator, but there are of course some effects that are more readily classified under this heading, such as improvement in facilities, support, spirit, applicants, integration, etc.

8.47 Effects on the Taxpayer.

These are partly considered in the section on costs below, but certain points are worth mentioning. We are using the term taxpayer and not rate-payer here to indicate a reference to the total tax structure, and the most important kinds of effects here are the possibility of very large-scale emulation of a given curriculum reform project, which in toto, especially with evaluation on the scale envisioned here, is likely to add a substantial amount to the overall tax burden. For the unmarried or childless taxpayer, this will be an effect which may with some grounds be considered a social injustice. Insofar as evaluation of a national armament program must be directly tied to questions of fair and unfair tax loads, the same must be applied in any national considerations of very large-scale curriculum reforms.

9. Values and Costs.

9.1 Range of Utility.

No evaluation of a teaching instrument can be considered complete without reference to the range of its applicability and the importance of improvement of education in that range. If we are particularly concerned with the underprivileged groups, then it will be a value of considerable importance if our new teaching instrument is especially well adapted for that group. It may not be very highly generalisable, but that may be offset by the social utility of the effects actually obtained. Similarly, the fact that the instrument is demonstrably usable by teachers with no extra training, sharply increases its short-term utility. Indeed it may be so important as to make it one of the goals of instrument development, for short-run high-yield improvements.

9.2 Moral Considerations.

Considerations of the kind that are normally referred to as moral have a place in the evaluation of new curricula. If the procedures for grading, or treating students in class, although pedagogically effective, are unjust, then we may have grounds for judging the instrument undesirable which are independent of any directly testable consequences. If one conceives of morality as a system of principles founded upon the maximising of extreme long-run social utility, based on an egalitarian axion, then moral evaluations should show up somewhere else on the criteria given above, as primary or secondary effects. But the time lag before they do so may be so long as to make it appropriate for us to introduce this as a separate category. There are a number of other features of teaching instruments that may be reacted to morally; 'the dehumanising influence of teaching machines' is a description often used by critics who are partly affected by moral considerations; whether misguidedly or not is another question. Curricula stressing the difference in performance on the standardised intelligence tests of negro and white children have been attacked as morally undesirable, and the same has been said of textbooks in which the role of the United States in world history has been viewed somewhat critically. Considerations like this will of course show up on a content-mediated approach to evaluation but they deserve a separate entry because the reaction is not to the truth or insight provided by the program, but to some other consequences of providing what may well be truths or insights, namely the consequences involving the welfare of the society as a whole.

9.3 Costs.

The costing of curriculum adoption is a rather poorly researched affair. Enthusiasts for new curricula tend to overlook a large number of secondary costs that arise, not only in the experimental situation, but in the event of large-scale adoption. Evaluation, particularly of items for purchase from public funds, has a strong committment to examination of the cost situation. Most of the appropriate analysis can be best obtained from an experienced industrial accountant, but it is perhaps worth mentioning here that even when the money has been provided for the salaries of curriculum-makers and field-testers and in-service training institutes there are a number of other costs that are not easily assessed, such as the costs of re-arrangements of curriculum, differential loads on other faculty, diminished availability for supervisory chores of the experimental staff (and in the long run, where the instrument requires more of the teacher's time than the one it replaces, this becomes a permanent cost), the 'costs' of extra demands on student time (presumably at the expense of other courses they might be taking), and of energy drain on the faculty as they acquire the necessary background and skills in the new curriculum, and so on through the list of other indirect effects many of which have cost considerations attached, whether the cost is in dollars or some other valuable.

10. Another kind of Evaluation - 'Explanatory Evaluation'.

Data relevant to the variables outlined in the preceding section are the basic elements for almost all types of evaluation. But sometimes, as was indicated in the first section, evaluation refers to interpretation or explanation in a different sense. While not considering this to be a primary or even a fully proper sense, it is clear from the literature that there is some tendency to extend the term in this direction. It seems to be preferable to distinguish between evaluation, and the attempt to discover an explanation of certain kinds of result, even when both are using the same data. Explanation-hunting is sometimes part of process research and sometimes part of other areas in the field of educational research. When we turn to considerations of this kind, data of a quite different variety is called for. We shall, for example, need to have information about specific skills and attitudes of the students who perform in a particular way, we shall call upon the assistance of experts who or tests which may be able to demonstrate that the failure of a particular teaching instrument is due to its use of an inappropriately advanced vocabulary, rather than to any lack of comprehensible organisation. Evaluation of this kind, however, is and should be secondary to evaluation of the kinds discussed previously, for the same reason that therapy is secondary to diagnosis.

11. Conclusions.

The aim of this paper has been to move one step further in the direction of an adequate methodology of curriculum evaluation. It is clear that taking this step involves considerable complication of the model of adequate evaluation study, by comparison with what has passed under this heading all too frequently in the past. Further analysis of the problem may reveal even greater difficulties that must be sorted out with an attendant increase in complexity. Complex experiments on the scale we have been discussing are very expensive in both time and effort. But it has been an important part of the argument of this paper that no substitutes will do. If we want to know the answers to the questions that matter about new teaching instruments, we have got to do an experiment which will yield those answers. The educational profession is suffering from a completely inappropriate conception of the cost scale for educational research. To develop a new automobile engine or a rocket engine is a very, very expensive business despite the extreme constancy in the properties of physical substances. When we are dealing with a teaching instrument such as a new curriculum or classroom procedure, with its extreme dependence upon highly variable operators and recipients, we must expect considerably more expense. The social payoff is enormously more important, and this society can, in the long run, afford the expense. At the moment its deficiency is trained manpower, so that short-term transition to the appropriate scale of investigation is possible only in rare cases. But the long-term transition must be made. We are dealing with something more important and more difficult to evaluate than an engine design, and we are attempting to get by with something like one percent of the cost of developing an engine design. The educational profession as a whole has a primary obligation to recognise the difficulty of good curriculum development, with its essential concomitant evaluation, and to begin a unified attack on the problem of financing the kind of improvement that may help us towards the goal of a few million enlightened citizens on the earth's surface, even at the expense of one on the surface of Mars.